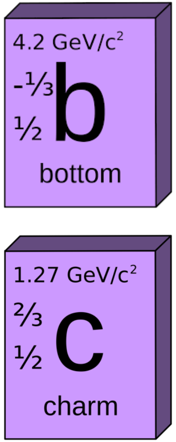
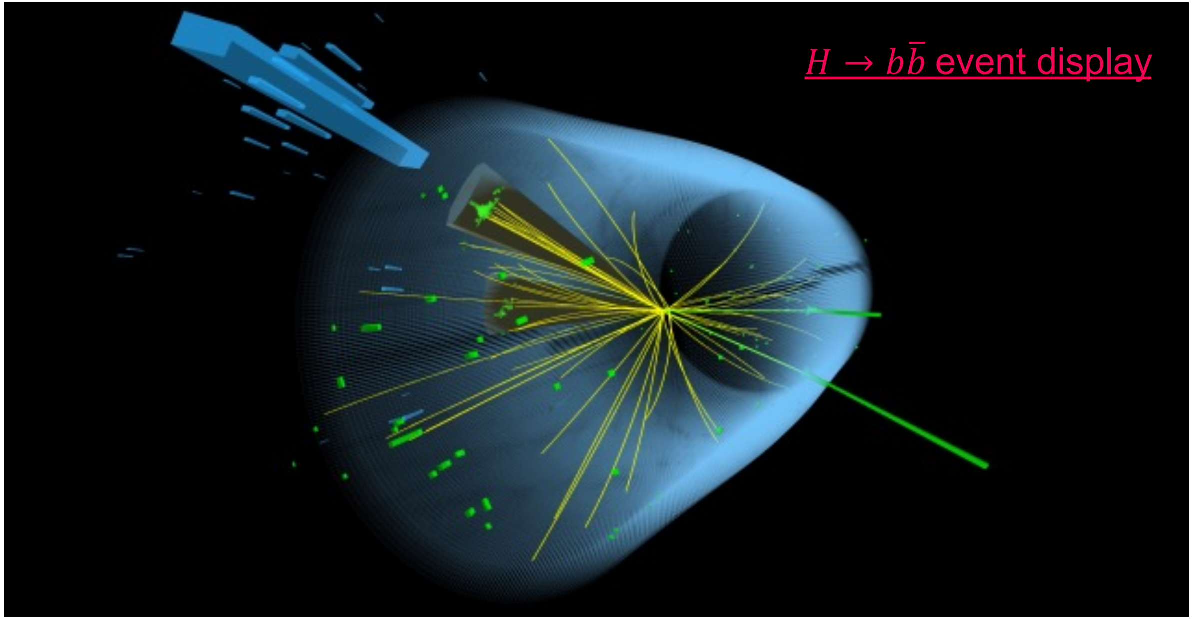
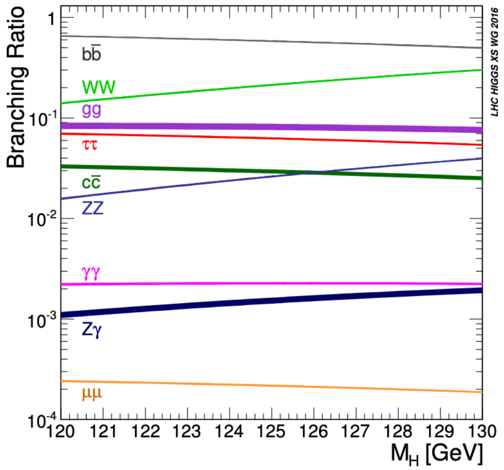
Heavy flavour tagging at the CMS experiment



Angela Zaza on behalf of the CMS collaboration
University of Bari & INFN

Motivation





R.L. Workman *et al.* (Particle Data Group), Prog. Theor. Exp. Phys. 2022, 083C01 (2022)

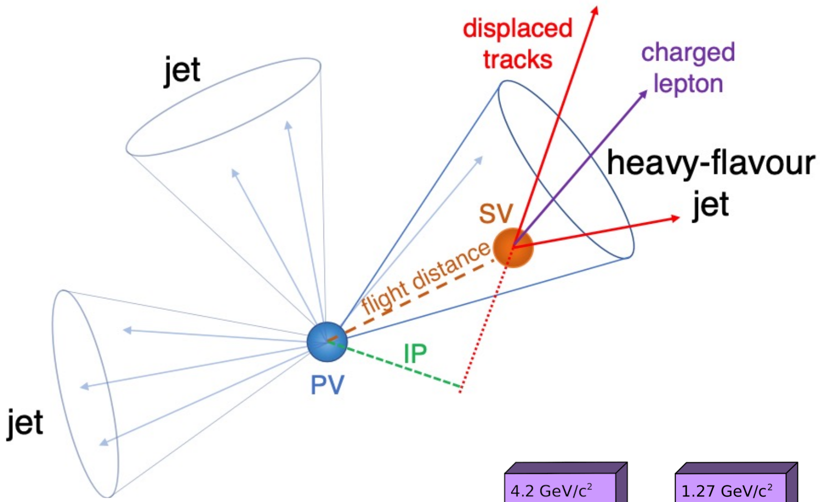
Classification of heavy flavor (b and c) jets crucial for several physics processes involving heavy quarks, such as **Higgs** decays

Heavy flavour jets

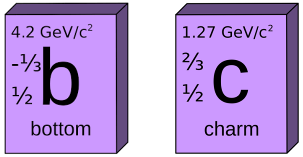
Jets originated from hadronization of b (c) quarks:

- ▷ Lifetime of b (c) hadrons ~ 1.5 ps (~1 ps)
  → **displaced tracks from PV** (impact parameter)
  → **SV**
- ▷ Larger mass and harder fragmentation w.r.t. light quarks and gluons
  → **larger p_T of the decay products**
- ▷ Presence of a **muon or electron** in 20% (10%) of the cases

*2018 JINST **13** P05011*



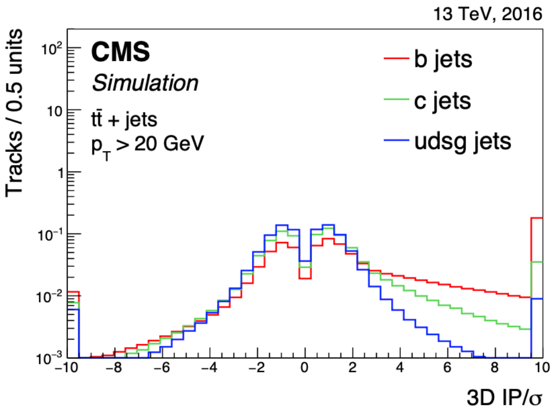
Heavy flavour tagging performed by combining many discriminating variables by means of MVA techniques

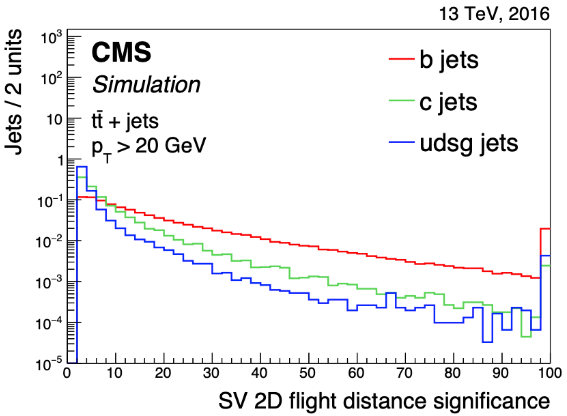


c-tagging more complex than b-tagging:
discriminating variable distributions intermediate between b and light-jet ones

Heavy flavour tagging: **DeepCSV**

- Deep Neural Network (**DNN**)
  4 hidden layers – 100 nodes
- **5 classes**:
  b (1 b), bb (2 b),
  c (1 c and no b), cc (2 c),
  lg (everything else)
- Jets reweighted to avoid $\mathbf{p_T}$ and $\boldsymbol{\eta}$ dependencies across flavours during the training
- Simulated samples used for training: $t\bar{t}$ and **QCD**





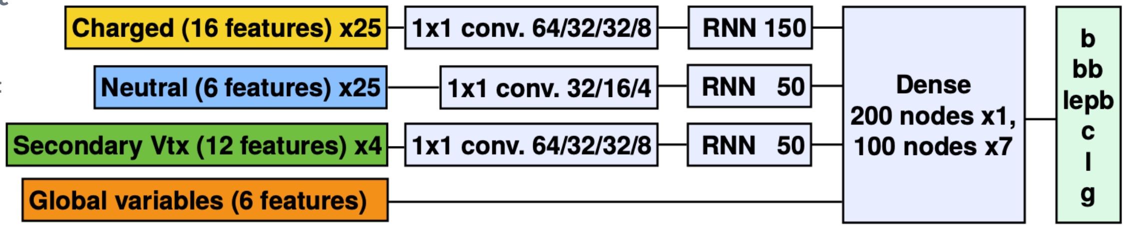
[2018 JINST **13** P05011](#)

Heavy flavour tagging:
DeepJet

[2020 *JINST* **15** P12012]

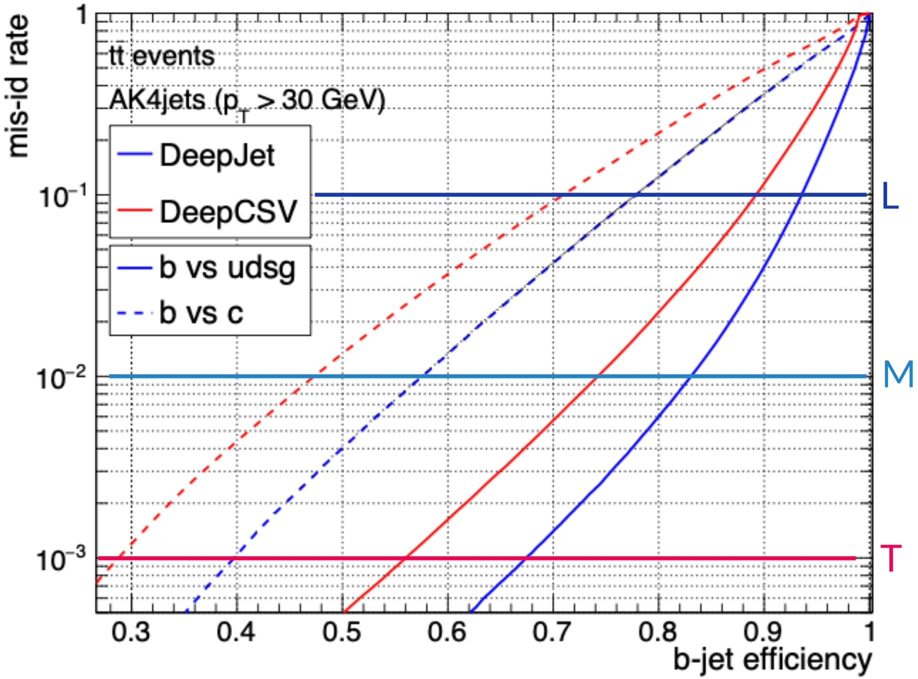
DNN, Convolutional NN (CNN) and Recurrent NN (RNN)

- ▷ Low level features from a large number of jet constituents
- ▷ Jets reweighted to avoid $\mathbf{p_T}$ and $\boldsymbol{\eta}$ dependencies across flavours during the training



- ▷ Automatic feature engineering performed for each constituent using 1x1 convolutional layers
- ▷ 3 RNN layers combine the information for each constituent sequence
- ▷ Fully connected layers combine the full jet and per-event level information
- ▷ **6 classes**:
  b, bb, **lepb** (leptonic b hadron decays)
  c, cc, **l** (uds), **g**
- ▷ Simulated samples used for training: $t\bar{t}$ and **QCD**

B-tagging performance



[2020 *JINST* **15** P12012]

$$P(BvsAll) = \frac{P(b) + P(bb) + P(lepb)}{P(b) + P(bb) + P(lepb) + P(c) + P(uds) + P(g)}$$

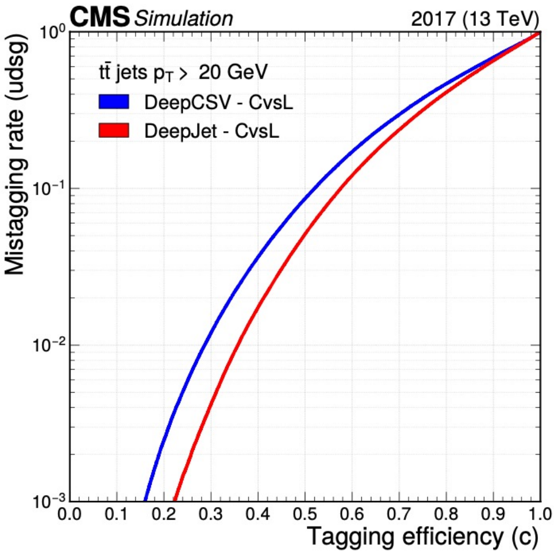
Working Points

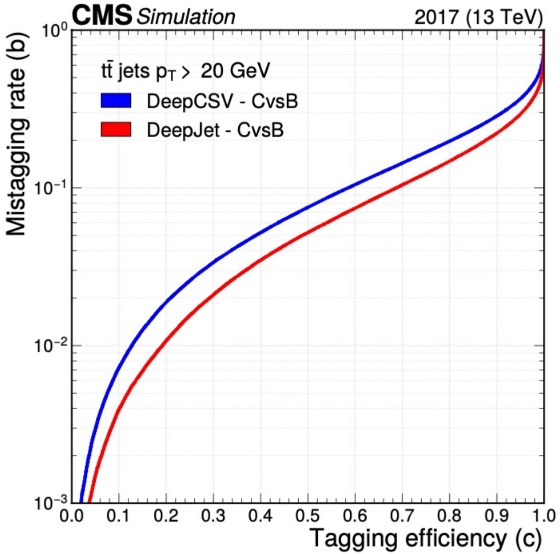
Loose (L): 10% udsg mis-id rate

Medium (M): 1% udsg mis-id rate

Tight (T): 0.1% udsg mis-id rate

C-tagging performance





[2022 JINST 17 P03014]

Two different scores for CvsL and CvsB discrimination

$$P(CvsB) = \frac{P(c)}{P(b) + P(bb) + P(lepb) + P(c)}$$

$$P(CvsL) = \frac{P(c)}{P(uds) + P(g) + P(c)}$$

Performance in data

- MC simulation does not provide a perfect representation of data $\rightarrow$ necessary to apply SFs to MC

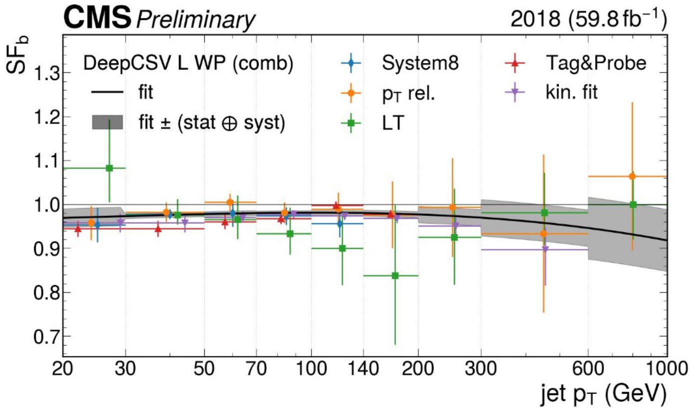
$$\varepsilon_f^{MC} = \frac{N_f^{Tagged}}{N_f^{Total}}$$

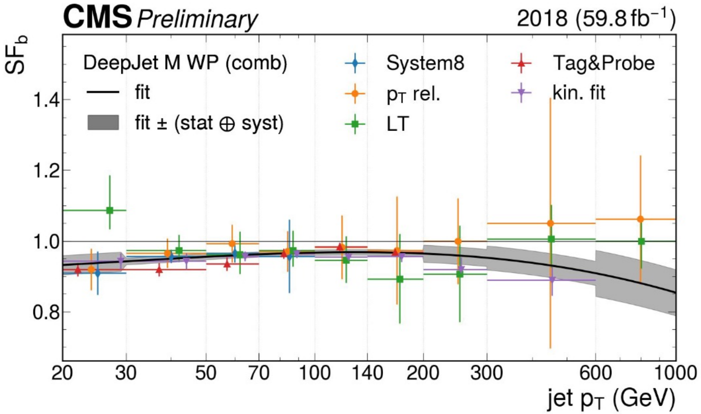
$$\varepsilon_f^{Data} = SF_f \times \varepsilon_f^{MC}$$

N_f^{Tagged}, N_f^{Total}, SF_f: number of tagged jets, number of total jets and calibration scale factor for the flavour f

- SFs calculated with different methods specific for QCD multijet, $t\bar{t}$, Drell-Yan and W+c events
- SFs evaluated at different WPs
- SFs also estimated as a function of the discriminator value with the IterativeFit method (crucial for analyses in which the full distribution of the b-tagging discriminating values is used, e.g. as inputs to an MVA)

b-tagging efficiency





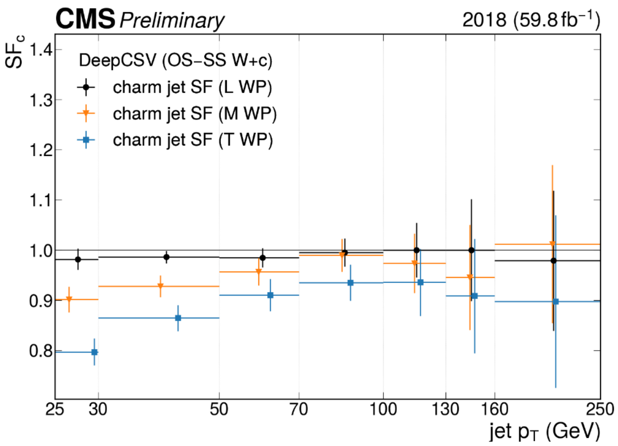
CMS DP-2023/005

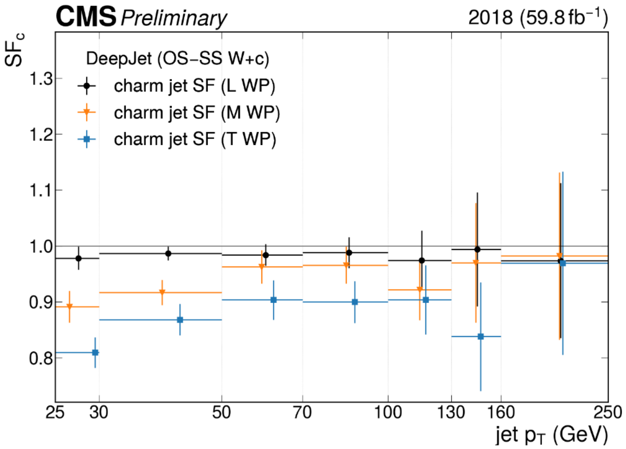
Performance in data:
c-tagging scores

- Three different sets of event selection, targeting W+c, $t\bar{t}$ and DY+jets/QCD events respectively c-, b- and light-enriched
- SFs calculated as function of both CvsL and CvsB

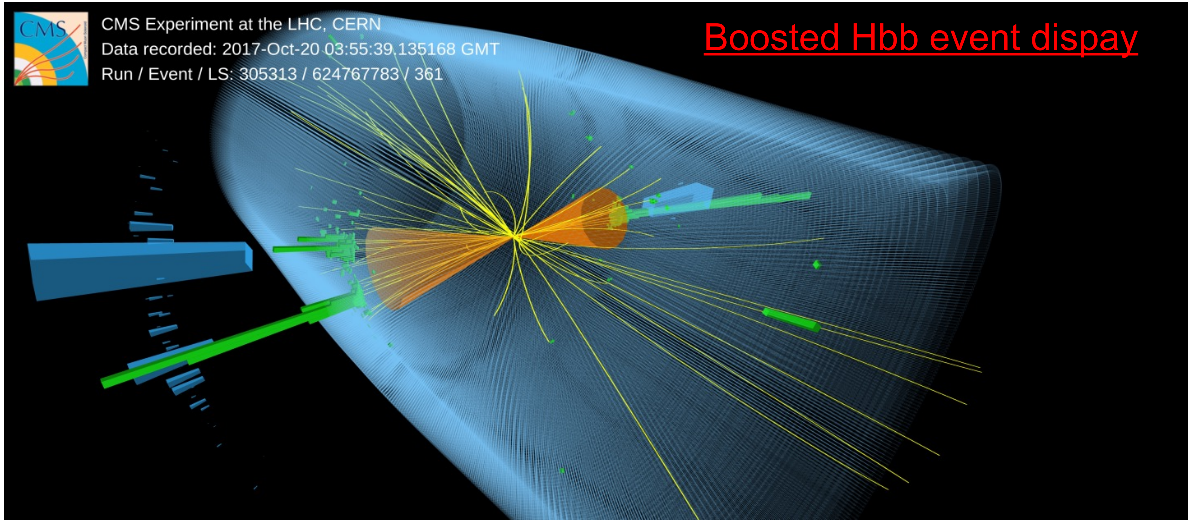
WP	DeepCSV					DeepJet				
	CvsL cut	CvsB cut	c eff.	b eff.	udsg eff.	CvsL cut	CvsB cut	c eff.	b eff.	udsg eff.
Loose	0.064	0.313	91.4%	35.0%	90.0%	0.038	0.246	94.4%	35.0%	90.0%
Medium	0.153	0.363	57.7%	25.0%	25.0%	0.099	0.325	63.7%	25.0%	25.0%
Tight	0.405	0.288	34.2%	20.0%	3.00%	0.282	0.267	40.3%	20.0%	3.00%

CMS DP-2023/006 JINST 17 (2022) P03014

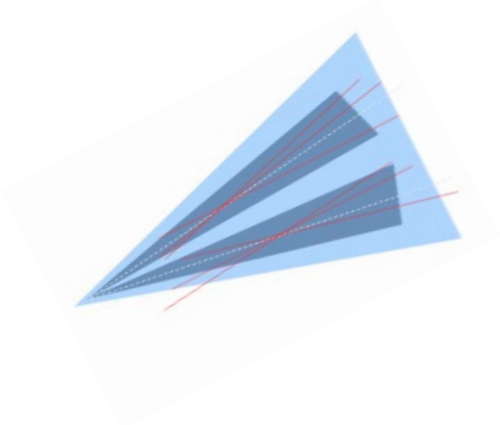




Heavy flavour tagging in **boosted topologies**



- At high energy, particles decaying to b or c quarks can be highly boosted and the decay products can result in **overlapping jets**



- In many analyses targeting $X \rightarrow q\bar{q}$ with $p_T^X \gg m_X$, large radius jets (i.e. AK8) are used

Double-b and **DeepDoubleX** taggers perform boosted jet ($q\bar{q}$) tagging

Heavy flavour tagging in **boosted topologies**

double-b tagger

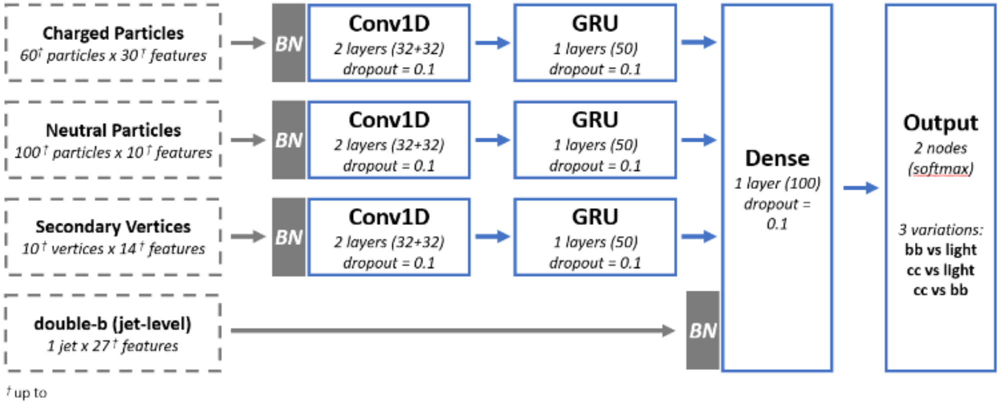
- Dedicated BDT algorithm for identification of the decay of a boosted object to a b quark pair
- 27 jet related properties exploited
- Input variables related to the correlation between the flight directions of the b quarks built by using the N-subjettiness axes to associate tracks and vetex to the subjets

The performance of the two taggers is evaluated by using AK8 jets ($\Delta R = 0.8$) in a boosted region of $300 < p_T < 1200$ GeV and mass window 20-200 GeV

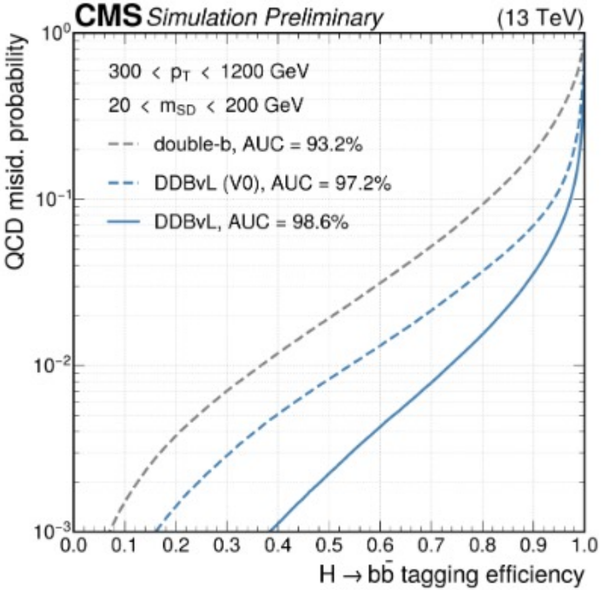
[CMS-DP-2022/041]

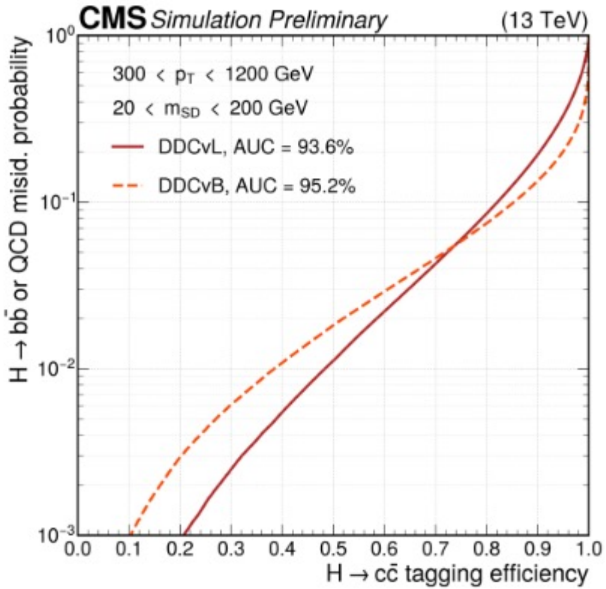
DeepDoubleX (DDX) tagger

- DNN algorithm for identification of the decay of a boosted object to a b or c quark pair
- Architecture and input variables set motivated by DeepJet
- Three separate taggers trained to distinguish H→ $b\bar{b}$ and H→ $c\bar{c}$ jets from QCD: **DDBvL**, **DDCvL**, **DDCvB**



Tagging performance in **boosted topologies**





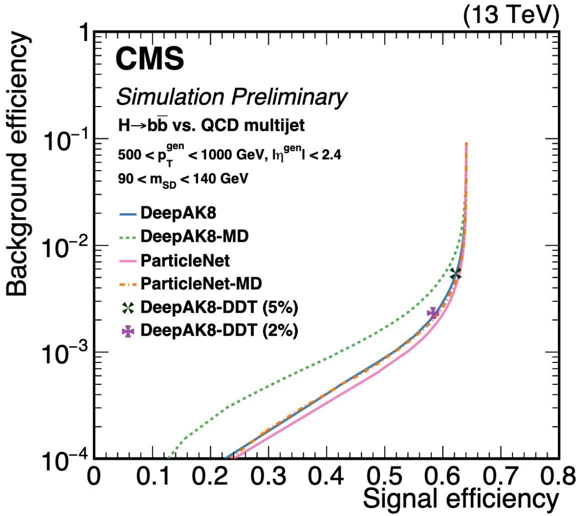
- MC simulation used for training and ROC estimation: QCD multijet, H→ $b\bar{b}$ and H→ $c\bar{c}$ events
- DeepDoubleX shows highly improved performance w.r.t. double-b tagger in H→ $b\bar{b}$ vs QCD discrimination

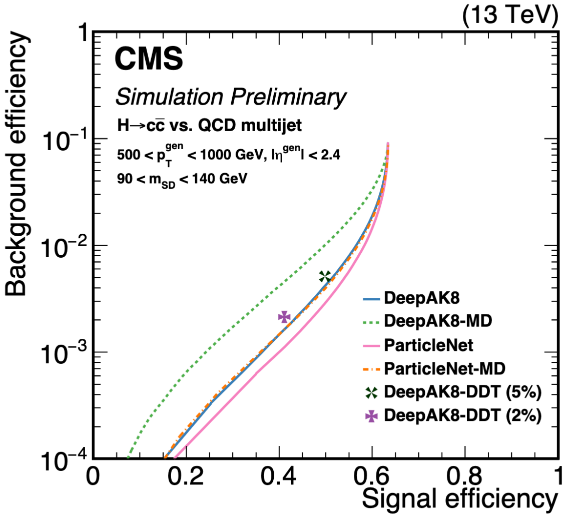
DDBvL (V0): earlier version of the DeepDoubleBvL
The latest version is mass-decorrelated by design (variable-mass Higgs MC samples are used) and exploits feature-ranking to prune and trim some input variables

Plans for Run3

ParticleNet

- Dynamic Graph CNN (DGCNN) considering jets as particle clouds
- Used for AK8 classification in some Run2 analyses (boosted Hbb/Hcc)
- Plans to use ParticleNet architecture for heavy flavour tagging during Run3





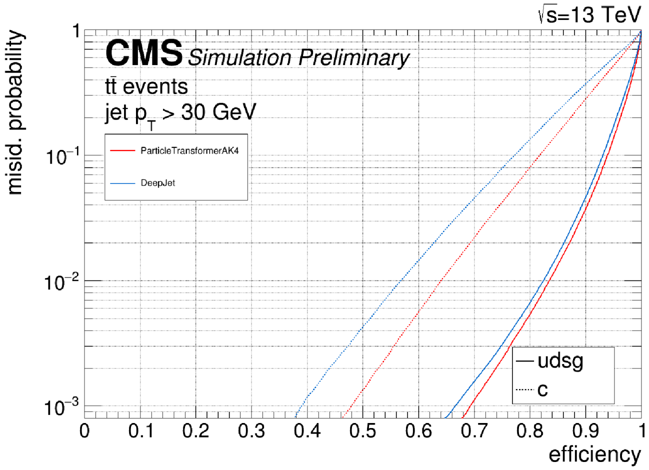
DeepAK8: previous DNN-based benchmark

CMS-DP-2020/002

Plans for Run3

ParticleTransformerAK4

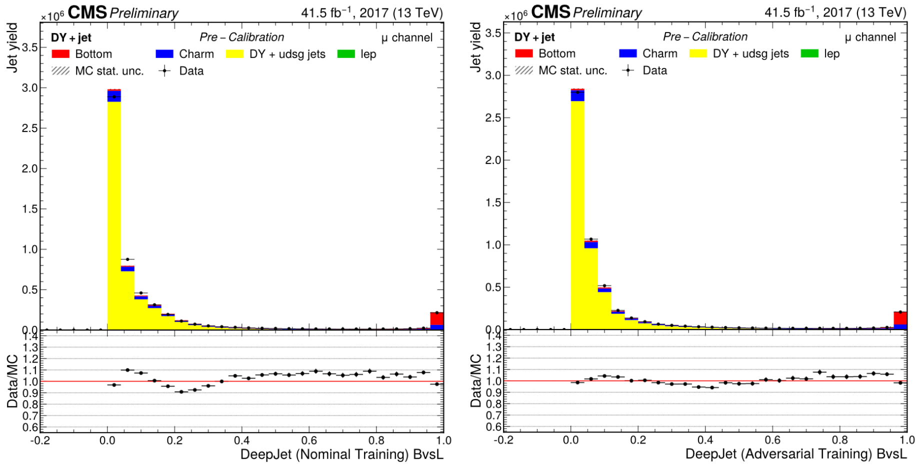
- Transformer neural network for AK4 jet tagging
- Additional input: pairwise «interaction» features between all jet constituent particles and secondary vertices



CMS-DP-2022/050

Adversarial training

- New strategy for reduction of data/MC differences prior to calibration and classifier robustness improvement
- Fast Gradient Sign Method (FGSM) attack used to systematically distort inputs



CMS-DP-2022/049

Summary

Heavy flavour tagging at CMS – Run2

- Comparison DeepJet/DeepCSV on AK4 jets
  - DeepJet shows much better performance
  - both taggers show good MC/data agreement
- Comparison double-b/DeepDoubleX on AK8 jets
  - DeepDoubleX outperforms and enables c-tagging

CMS physics programme largely benefits from these powerful tagging algorithms

New strategies for Run3

- ParticleTransformer and Adversarial training

Thank you for listening!

angela.zaza@cern.ch

Back-up

angela.zaza@cern.ch

DeepAK8

DeepAK8: multi-class particle identification algorithm for identifying hadronic decays of highly Lorentz-boosted top quarks and W, Z, and Higgs bosons for AK8 jets

Two lists of inputs defined for each jet:

1. **Particle list**: up to 100 jet constituent particles, sorted by decreasing pT. Measured properties (42) of each particle (p_T, energy deposit, charge, angular separation between the particle and the jet axis, etc) For charged particles, additional information measured by the tracking detector is also included.
2. **SV list**: up to 7 SVs, each with 15 features, such as the SV kinematics, the displacement, and quality criteria.

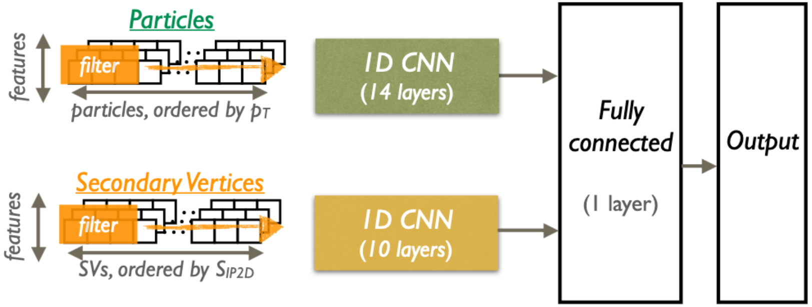


Figure 9: The network architecture of DeepAK8.

CMS-PAS-JME-18-002

Soft-Drop (SD)

Algorithm that recursively removes wide-angle radiation from a jet.

It depends on two parameters, a soft threshold z_{cut} and an angular exponent β.

1. Break the jet j into two subjets by undoing the last stage of C/A clustering. Label the resulting two subjets as j_1 and j_2
2. If the subjets pass the condition $\frac{\min(p_{T1}, p_{T2})}{p_{T1}+p_{T2}} > z_{\mathrm{cut}} \left(\frac{\Delta R_{12}}{R_0}\right)^{\beta}$, j is the final soft-drop jet
3. Otherwise, redifine j to be equal to subjet with the larger p_T

JHEP **05** (2014) 146

N-subjettiness

Jet shape variable, computed under the assumption that the jet has N subjets, and it is defined as the p_T -weighted distance between each jet constituent and its nearest subjet axis (ΔR):

$$\tau_{\mathrm{N}} = \frac{1}{d_0} \sum_{k} p_{\mathrm{T}}^{k} \min(\Delta R_{1,k}, \ldots, \Delta R_{\mathrm{N},k}),$$

2018 JINST ***13*** *P05011*

Where *k* runs over all jet constituents. The τ_N variable has a small value if the jet is consistent with having N or fewer subjets. The subjet axes are used as a starting point for the τ_N minimization. After the minimization, the τ_N axes, also called τ axes, are obtained.

Fast Gradient Sign Method (FGSM) attack

It is used to systematically distort inputs based on the geometry of the loss surface and acts on inputs x_{raw} as follows:

$$x_{\mathbf{FGSM}} = x_{\mathbf{raw}} + \epsilon \cdot \mathrm{sgn}\left(\nabla_{x_{\mathbf{raw}}} J(x_{\mathbf{raw}}, y)\right)$$

CMS-DP-2022/049

where y refers to truth labels, and J is the loss function.